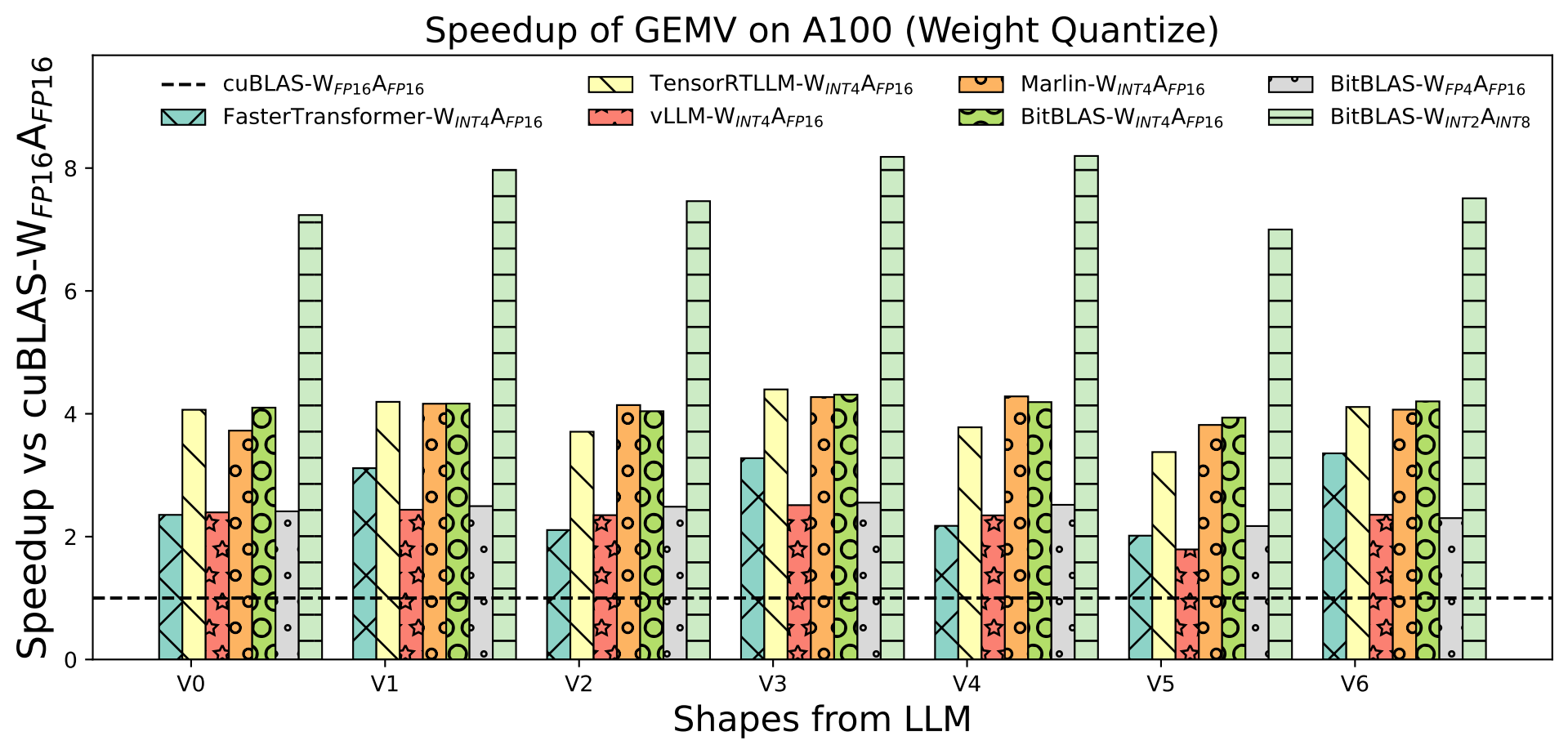
Speedup of GEMV on A100 (Weight Quantize)
cuBLAS-W$_{FP16}$A$_{FP16}$
FasterTransformer-W$_{INT4}$A$_{FP16}$
TensorRTLLM-W$_{INT4}$A$_{FP16}$
vLLM-W$_{INT4}$A$_{FP16}$
Marlin-W$_{INT4}$A$_{FP16}$
BitBLAS-W$_{INT4}$A$_{FP16}$
BitBLAS-W$_{FP4}$A$_{FP16}$
BitBLAS-W$_{INT2}$A$_{INT8}$
Speedup vs cuBLAS-W$_{FP16}$A$_{FP16}$
0
2
4
6
8
V0
V1
V2
V3
V4
V5
V6
Shapes from LLM